



This document is classified as *FORMAT* Public

EVALUATION METRICS

by
Katrien Vanherck (PNO)

Abstract

This document describes the evaluation metrics that will be used to assess the FORMAT methodology performance during and after the FORMAT project, from the perspective of the methodology builders (the FORMAT team that worked on WP3), the methodology users (e.g., the core team using the FORMAT methodology during the test cases) and the beneficiaries of the outcome (generally: the decision-maker who will benefit from the results of the forecast). The document will describe the way the evaluation metrics were conceived out of the discussions and literature sources.

Approval status

Author:	Katrien Vanherck	PNO	ER, seconded to PoliMI
Approved by:	Gaetano Cascini	PoliMI	Project Coordinator

Dissemination

PU

Document details

Issue Date:	30.10.2013
Contract n°:	PIAP-GA-2011-286305
Project n°:	286305

Revision table

Issue	Issue date	Modifications
0.1	21.10.2013	Draft
1.0	25.10.2013	Final issue

Electronic file details

Master file location	FORMAT collaborative space
Filename	FORMAT_deliverable 4.1.doc

0. Index

0. Index	3
1. Why we need evaluation metrics.....	4
2. Process of developing the metrics	5
2.1. Preliminary evaluation questions	5
2.2. Comparison with literature	6
2.3. Evaluation of the learning process.....	8
3. Input for Gate questions and Handbook	9
3.1. Input for Gate Questions.....	9
3.2. Input for handbook.....	11
4. Final evaluation metrics	12
4.1. Evaluation matrix	12
4.2. Quantifying the value of the forecast output/outcome.....	12
5. Bibliography.....	13
Annex I: Evaluation Matrix	14
Annex II: Scoring card for beneficiaries	17

1. Why we need evaluation metrics

According to the tasks in WP4 of the FORMAT project, the goal of the evaluation metrics is described as follows:

T4.1 Definition of an evaluation metric for the testing activities (led by PoliMI).

Definition of a metric for the **evaluation of the benefits and the efforts** associated to the proposed methodological approach. These metrics will be co-defined by the partners involved in the project and will be applied to the Pilot Projects implemented during the second and third year.

T4.3 Small project: quality assessment of the proposed methodology and value estimation of the products of the test cases (led by PoliMI).

Methods developed within the FORMAT project will undergo a careful validation and review activity: the second year will be dedicated to testing the proposed roadmap by means of two complementary case studies and the validation criteria defined in Task 4.1 will be **applied to validate and tune the guidelines developed in WP3**.

T4.5 Extended projects: quality assessment of the proposed methodology and value estimation of the products of the test cases (led by PoliMI).

Methods developed within the FORMAT project and reviewed after the second year testing activity will undergo a further validation and demonstration: in this Task, still using the same evaluation metrics defined in Task 4.1, the **overall methodology and the IT Demonstrator will be analyzed and critically discussed**.

Initially, the discussions were focused on how to analyse the performance of the forecast: are the results reliable? Since the results forecast a situation in the future, the reliability can -in fact- only be really checked after 3-5 years, at least, depending on the time span of the forecast. Therefore, a starting document was prepared by PoliMI (Politecnico di Milano, Milan, Italy) and WHRIT (Whirlpool Italy), with ideas and suggestions for indicators that could analyse the performance of a forecast already during the FORMAT project. This mostly involved analysing the effort-to-benefit ratio (are the efforts acceptable for the new knowledge that we gained?) and the non-obviousness of the results (could I have arrived to this conclusion without performing the forecast?).

With the input of PNO, at the start of the secondment of the author, it has been decided that the evaluation metrics should not be strictly limited to the assessment of the forecast results (the content of the output of the methodology), but also include the evaluation of the methodology itself (Does it work properly? Are the methods feasible? Is the output presentable?). This will be needed in order to complete task 4.3 and 4.5, where the FORMAT methodology needs to be assessed before finalizing the methods as well as the guidelines. This evaluation should also deliver inputs in terms of how to present the outputs of the forecast, how to prepare the forecast report and how to prepare the guidelines and manual.

At the ToK (Transfer of Knowledge) meeting at WHRIT (17-18 Sept), a general consensus was reached on the structure of the evaluation metrics:

- The metrics will need to be useful to evaluate two related but distinct topics:
 - The **forecast outcome** (content of the output)
 - The **methodology and methods** (efforts, exhaustiveness, resources, presentation of output, ...)
- And the metrics will be formulated for three kinds of evaluators, in order to get the satisfaction from different perspectives:
 - The **decision maker** (mainly the content and presentation of the output)
 - The **user** of the methodology (is it feasible, what expertise is needed, is it easy to interpret the results, is it manageable to assemble the report, ...)
 - The **builders** of the methodology (assessing the usefulness of and the resources/efforts needed for the different methods during the test cases)

2. Process of developing the metrics

The evaluation metrics were finalized after different development stages. At the first stage, preliminary evaluation questions were more intuitive (compared to the final questions) and not yet ordered into different categories. In a second stage, the preliminary metrics were compared with the best practices in literature and enhanced by adding questions that should be incorporated into the methodology itself, rather than being used as an evaluation question outside of the process. In a third stage, the questions were reduced and structured into an evaluation matrix.

2.1. Preliminary evaluation questions

A first proposal for evaluation questions, to be answered positively or negatively, can be found in the table below. These metrics were discussed with the Consortium partners during the ToK meeting at WHRIT in September 2013.

Some comments were made regarding the usability of the questions (e.g. “Will it ever be possible to compare the outcome of FORMAT to the outcome of other methodologies with the same data input?”). The practical implementation of other questions was addressed (e.g. “Will different research groups using the FORMAT methodology with the same data get to the same result?” can be formed through student projects at PoliMI).

PRELIMINARY STRUCTURE BEFORE TOK MEETING	
Evaluation of the forecast results (in charge of the recipients of the forecast)	
	<i>Usefulness of the forecast results</i>
	Are they adequate to analyse the impact on <ul style="list-style-type: none"> - The technical performance of the analysed technology? - The economic performance of the analysed technology? - The sustainability of the analysed technology?
	Are they useful for decision-making?
	<i>Non-obviousness of the forecast results</i>
	Could you have reached these conclusions without doing the forecast?
	Is the result unexpected?
	Is the result acceptable (likely to be correct)?
	<i>Value of the forecast results</i>
	How much do you value the result in terms of usefulness and non-obviousness?
	Are there added benefits associated with the forecast output? <ul style="list-style-type: none"> - Growth in knowledge or skills? - Useful for other purposes (e.g. marketing)
	Quantify the efforts and weight against the benefits <ul style="list-style-type: none"> - Time - Money/information - Capacity/HR
Evaluation of the methodology as a whole (in charge of the applicants of the methodology)	
	<i>Did we reach the objectives?</i>
	Make list of 'generic' types of objectives and translate to relevant indicators
	<i>Collectively exhaustive and mutually inclusive</i>
	Do we need each step?
	Is any step missing?
	Do we need complementary data sources?
	<i>Are the efforts acceptable?</i>
	Do we stay within requested limits of time/budget/capacity?
	Is the necessary expertise available?
	Is the training effort acceptable?
	<i>Is the methodology replicable?</i>
	Do two teams reach the same result for the same situation?
	Are the results comparable with other methodologies?

Evaluation of each of the methods (in charge of the applicants of the methodology)	
	<i>Is the method useful?</i>
	Does it contribute to the objectives?
	In how many test cases was it used?
	<i>Is the method user-friendly?</i>
	Beginner level or advanced?
	Easy-to-learn?
	Easy-to-interpret?
	<i>How many resources are required?</i>
	Time, capacity, budget
	Expertise, information
	<i>Is an alternative method available?</i>
	For which expertise level?
	With higher or lower use of resources?
	Clear when to choose which method?
	<i>Is the output suitable?</i>
	Easy-to-interpret?
	Must-have or not necessary to show in the report?
	Qualitative or quantitative?
	Attractive visualization?

2.2. Comparison with literature

A chapter in the book “Principles of Forecasting” by Armstrong (2002) on “Evaluating methods” was consulted in order to complete and refine the evaluation metrics for the FORMAT methodology.

Overall, Armstrong states that the evaluation of forecasting methods should comprise the following 4 parts:

- 1) Testing assumptions;
- 2) Testing data and methods;
- 3) Replicating outputs;
- 4) Assessing outputs.

The entire first step, ‘testing assumptions’, should actually be incorporated into the FORMAT methodology in Stage 5, where the validity of the forecast will be assessed. The second and third step, ‘testing data and methods’ and ‘replicating outputs’ are largely in accordance with the preliminary metrics for the builders and the users of the forecast. The fourth step, ‘assessing outputs’, should match the evaluation metrics that are useful for the decision makers.

The specific evaluation methods for each of these stages, which are provided by Armstrong, were studied more in detail in order to see which ones could be most relevant to include as metrics in the FORMAT project.

The table below lists some suggested evaluation methods and/or metrics indicated by Armstrong (more elaborate explanation can be found in the reference) and where they are already included or where they could still be included. These suggestions were discussed more in detail in order to:

- Decide upon whether to include or exclude extra methods and if so, how and where;
- Reduce the number of evaluation questions to make the process of evaluation less tedious.

Suggested by Armstrong	Already included in	To be used by
Use objective tests of assumptions	Stage 5	Users
Test assumptions for construct validity	Stage 5	Users
Describe conditions of forecasting problem	Stage 1-2	Users
Design forecasting tests to match the forecasting problem	Partially taken into account during the selection of relevant test cases	Users
Tailor analysis to the decision	Stage 1-2, Stage 6	Users
	Evaluation metrics on usefulness	Beneficiaries
Assessing the reliability and validity of the forecast	Stage 5	Users
	Evaluation metrics on methodology outcome	Beneficiaries
Disclose details of the methods	Stage 6	Users
	FORMAT handbook	
Find out whether the client understands the method	Evaluation metrics on individual methods	Builders/Users
Use direct replications to identify mistake	Evaluation metrics on methodology (project in class for students by Dmitry)	Users
Replicate studies to assess their reliability (use same data with different methods or different data with same methods)	Carrying out different test cases to assess the methodology	Users
	Evaluation metrics for methodology	
Compare forecasts obtained by different methods	Evaluation metrics for methodology	Users
Examine all important criteria	Evaluation metrics for the outcome	Beneficiaries
Assess face validity (is it reasonable)	Evaluation metrics for the outcome	Beneficiaries
Explicit cost-benefit analysis	Evaluation metrics for the outcome, methodology and individual methods	Users/Beneficiaries
Armstrong method	Could still be included in	To be used by
Compare a method to its reasonable alternatives	The evaluation of methods in each stage of the FORMAT methodology	Builders
Design forecasting tests to match the forecasting problem (+ compare with backcasting scenarios)	The evaluation of the methodology and its outcome	Builders (e.g. add as another project in class for students at PoliMI)
Provide easy access to data	The evaluation of the outcome	Users
Describe potential sources of bias by forecasters	To be included in Stage 5 in the methodology	Users
Extend studies to assess their generalizability	The evaluation of the outcome and the methodology	Users Builders
Conduct extensions in realistic situations	The evaluation of the outcome of the methodology	Builders
Prespecify criteria.	Stage 1-2	Users Beneficiaries
Error measures <ul style="list-style-type: none"> - Not affected by scale - Valid - Avoid high sensitivity to degree of difficulty - Avoid bias - Avoid sensitivity to outliers - Do not use R-square to compare forecasting models (time series) - Do not use Root Mean Square Error for comparisons across series - Use multiple error measures - Ex-ante and ex-post tests - Statistical significance only to compare accuracy - Large sample of independent forecast errors 	Stage 4	Builders Users

2.3. Evaluation of the learning process

Carrying out a forecasting project should be considered in itself a learning process that can provide, next to the immediate forecasting output, other benefits to the users and beneficiaries. This is mainly associated with their growth in knowledge and skills. Even the 'builders' of a forecasting methodology, or more in general the experts that conduct research on forecasting methodologies, will learn new things from each application that they are involved in and broaden their knowledge purely by applying the methodology to different types of forecasting cases.

The evaluation should therefore consider also what can be learned through carrying out a forecasting project from the three perspectives, also considering how this knowledge can be transferred to other practices (of the forecasting expert / in the company).

Scores:

- Poor
- Fair
- Good
- Excellent

For users:

How much did your knowledge or skills improve in terms of:

- Working with systematic step-by-step processes and methodologies
- Project management
- Team work
- Handling IP data, e.g. patent searching, ...
- ICT, e.g. text mining, ...
- The position of your company in the market, in terms of R&D leadership, ...
- The external factors that are related to the processes in your company
- ...

For beneficiaries:

How much do you think you / your team improved in terms of:

- (Same as listed for users)

For builders:

How much did your knowledge or skills improve in terms of:

- The strengths and weaknesses of different forecasting methods
- The way of combining different forecasting methods
- The interaction with company experts and how to explicit their tacit knowledge
- The amount of training needed for non-experts to be able to contribute to a forecasting project?
- The practical applicability of different methods in a company environment
- ...

3. Input for Gate questions and Handbook

A further refinement of the evaluation questions led to three separate lists of questions:

- A list of questions that can be used to define the Gate requirements;
- A list of questions that can be used to define the Guidelines in the Handbook.
- An evaluation matrix to be used for the evaluation of the FORMAT methodology and its outcome, during the FORMAT project by the FORMAT team (evaluating the test case results) as well as after the FORMAT project, once the methodology is being used in real case scenario's, by the users and beneficiaries.

3.1. Input for Gate Questions

In order to assess whether the stage has provided all required information (at sufficient quality) that is needed to pass the gate, the following Gate Questions could be included into the FORMAT methodology at each Gate checkpoint.

This is not an exhaustive list. There may still be other questions that need to be asked at each gate. For example, stage 5 is not solely dedicated to the validation of the forecast result. In fact, the first and main objective of this stage is to integrate all the collected information, interpret the results and translate them into a coherent forecast study. Some evaluation questions need to be added to evaluate this process (e.g. bias in interpretation, consideration of different scenarios,...)

POTENTIAL GATE QUESTIONS	
Stage 1	
1.1.	Did we describe the forecast scope?
1.2.	Did we describe the time period of data and information to take into account?
Stage 2	
2.1.	Do we know which questions need to be answered?
2.2.	Do we know which decisions need to be made?
2.3.	Do we know within which limits the forecast needs to be completed?
	2.3.1. How many people can be involved internally and externally?
	2.3.2. For how long?
	2.3.3. When do we need the results?
	2.3.4. What budget can be spent?
2.4.	Do we know the sources of data and information?
Stage 3	
3.1.	Did we choose the best method from the possible alternatives?
	3.1.1. Could we have reached the same data/information with less effort?
3.2.	Did we choose the right boundaries in comparison to the drivers and barriers?
3.3.	How much effort did we spend in this stage?
	3.3.1. How long did the stage take to complete?
	3.3.2. How many person months were spent?
	3.3.3. How much budget did we spend?
Stage 4 (for each substage)	
4.1.	Did we choose the best method from the possible alternatives?
	4.1.1. Could we have reached the same data/information with less effort?
4.2.	How much effort did we spend in this stage?
	4.2.1. How long did the stage take to complete?
	4.2.2. How many person months were spent?
	4.2.3. How much budget did we spend?
Stage 5	
5.1.	Are the assumptions reasonable and valid (by using an objective test)?
5.2.	What are the limits of validity of the forecast?
5.3.	Are there potential sources of bias from the users?
	5.3.1. How do you rate your expertise in the field of ...? (before starting)
	5.3.2. How do you rate your expertise & contribution to the forecast in the field of... (after finishing)?
	5.3.3. Are there any discrepancies between 5.3.1. and 5.3.2.?
5.4.	How much effort did we spend in this stage?

	5.4.1. How long did the stage take to complete?
	5.4.2. How many person months were spent?
	5.4.3. How much budget did we spend?
Stage 6	
6.1.	Did we answer the required questions? (See 2.1)
6.2.	Did we provide recommendations useful for the decisions that need to be made? (See 2.2.)
6.3.	Did we follow the project plan and stay within the required budgetary and time limits? If no, then why not? (See also 3.3., 4.2., 5.4.)
6.4.	Did we spend a reasonable amount of efforts in each stage? Were the efforts balanced with the output in each stage?
6.5.	Is the output presented in an attractive and meaningful way to the beneficiary?

3.2. Input for handbook

A self-assessment feature should become a part of the FORMAT Handbook. Users of the FORMAT methodology can then assess themselves

- **before starting** the forecasting project, to see whether they are ready (i.e. they have the competences) to start using the methodology;
- **after completing** the forecasting project, whether they followed and carried out the methodology according to all requirements, whether they were in fact missing some expertise in order to get reliable results, whether they have spent more or less a balanced amount of time on each of the stages,

Self-assessment information can be collected during the testing phase of the FORMAT methodology in order to assemble this chapter. We can include information such as:

- A general **time-slot division** for the different stages in the methodology (e.g. pie diagram: in general stage 1 should only take 1%, stage 2 maximum 5%, stage 3 maximum 20%, stage 4 maximum 60%, etc... of the total time allocated to the forecast;
- An **estimate of the required efforts, expertise, sources of information and data** that are associated with different methods that can be used in a certain stage, in order to support users in selecting the most appropriate method for a particular situation;
- A general **“read this first” chapter** that describes the minimal expertise and training that a user should have (completed) in order to be able to fully execute the methodology in a proper manner;
- A **stage-by-stage “test-yourself” feature** that shows users whether they are ready to start this stage of the methodology;
- ...

The following questions can provide the necessary input in order to assemble such chapters in the FORMAT handbook, e.g. for a **“method selection guide”** in each of the stages where the users are allowed to choose between a variety of methods in order to get the information required at the gate:

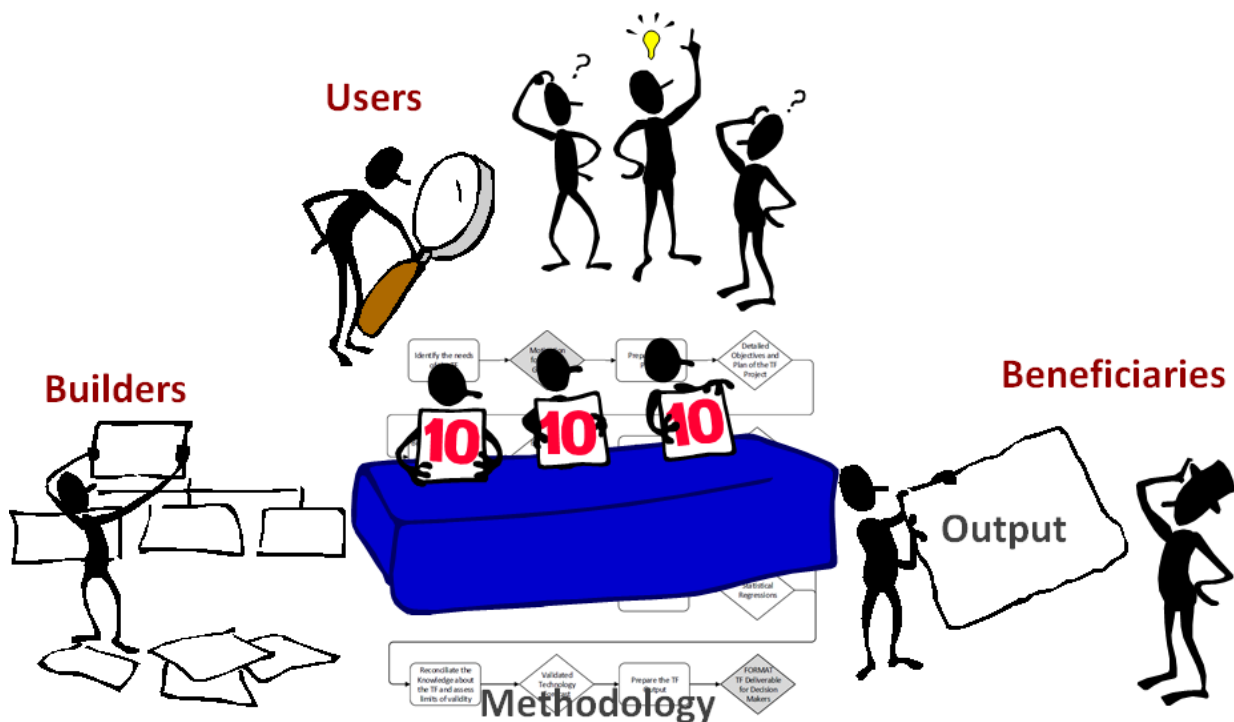
- What is the expertise that a user should have in order to use a method?
- How fast can a non-expert learn the method?
- Can the results be easily interpreted?
- What is the required effort in terms of time, person months, budget?
- At the minimum, what sources of information or data may be needed as input?
- What alternative methods are available?
- Is it clear which method is more appropriate for different situations (e.g. available time, available expertise, available information or data)
- Is the output of this method easy to interpret? (input from user)
- Can the output be presented in a meaningful way to the beneficiary? (input from beneficiary)
- Is the output of this method a must-have in the final report of the results? (input from user and beneficiary)

4. Final evaluation metrics

The final evaluation metrics are built around the principle as discussed in the introduction, namely:

- evaluating two distinct topics (the methodology and its outcome)
- from three different perspectives (the builders, users and beneficiaries).

This is illustrated in the following figure:



4.1. Evaluation matrix

A **matrix** with the different evaluation questions is provided as **Annex I** to this deliverable. The questions are structured according to:

- What is being evaluate;
- Who should answer the questions.

4.2. Quantifying the value of the forecast output/outcome

The metrics concerning the outcome of the methodology are qualitative assessments of the beneficiaries. These assessments will be made quantifiable by asking to choose between 4 qualitative levels that are each given a statistically relevant score, where

- level 1 is “definitely no” (scoring 2);
- level 2 is “probably no” (scoring 5);
- level 3 is “probably yes” (scoring 7);
- level 4 is “definitely yes” (scoring 10).

The scores will be counted in a statistically appropriate way to measure how valuable the output of the forecast is to the beneficiary.

The scorecard is provided as **Annex II**.

5. Bibliography

- [1] “Principles of Forecasting” by Armstrong (2002) on “Evaluating methods”

Annex I: Evaluation Matrix

	BUILDERS	USERS	BENEFICIARIES
PRACTICAL USE OF METHODS OR METHODOLOGY	How much effort did we spend in the whole process? - Time? - Person months? - Budget?	How much effort did we spend in the whole process? - Time? - Person months? - Budget?	
	Which different types of expertise are required to carry out the whole methodology?	Which different types of expertise are required to carry out the whole methodology?	
	Are the efforts spent in each stage balanced towards the relevance and importance of the output of the stage?		
		How much training did you need in order to use the methodology?	
		Did the core team change during the project? (Did they remain motivated to continue? Did they find the experience useful or interesting?)	
			Are the efforts needed to complete the forecasting project acceptable?
DETECTING GAPS, INCONSISTENCIES, DRAWBACKS	Did each method contribute to at least one objective?		
	Could we have arrived to the conclusions of any stage without using the suggested methods?	Could we have arrived to the conclusions of any stage without using the suggested methods?	
	After each stage, do we have all that is needed to be able to carry out the next stage or is there any data or information missing?	After each stage, do we have all that is needed to be able to carry out the next stage or is there any data or information missing?	
	Are the methods and stages balanced in terms of content (scientific domain)?		
	Are the methods and stages balanced in terms of qualitative vs. quantitative data and information?		

	Is the reasoning behind the interpretation correct or did we make deduction errors?		
	Can we think of any relevant questions that will not be answered by use of the methodology?		
	Can the methodology be extended to industries and technologies other than manufacturing?		
	Compared to the state of the art (Del. 2.3 and 2.5), what are the improvements in the methodology?		
		Are there any methods that you found to be not useful?	
		Are there any methods that you found to be too difficult (e.g. hard to interpret the results)?	
		In the stages were more than one method is allowed, which method did you prefer?	
			Does the methodology overlook any important factors?
VALUE OF OUTCOME (scaling 1=more 'no' to 4=more 'yes')			Were the stated objectives of the project appropriately addressed by the methodology?
			Does the output of the methodology provide an answer to the required questions?
			Are the results adequate to analyse the impact on the technical performance of the analysed technology?
			Are the results adequate to analyse the impact on the economic performance of the analysed technology?
			Are the results adequate to analyse the impact on the sustainability of the analysed technology?
			Do you think that you would not have arrived to these conclusions without doing the forecast?
			Is the result unexpected or surprising in some way?
			Can you accept that the result is likely to be correct?

			Are the recommendations based on the output useful for decision-making?
			How much do you value the result in terms of usefulness and non-obviousness?
			Did the forecast help to gain new insights about the environment of the company?
			Did the forecast help you to reduce uncertainties in projects or strategies...or to adapt to such uncertainties?
			Did the forecast lead to a better understanding of the market (potential customers, customer needs, competition)?
			Does the forecast improve the capability of the company to accept or adopt alternative perspectives of the future?
REPLICABILITY	Do different teams reach the same results based on the same data and information? (= student project)		

Annex II: Scoring card for beneficiaries

Please indicate your satisfaction with the forecast results by scoring the following questions:				
Tick only one box, where: 1 = definitely no 2 = more no 3 = more yes 4 = definitely yes				
	1	2	3	4
Scoring:	2	5	7	10
Were the objectives set for the project appropriately addressed by the methodology?				
Does the output of the methodology provide an answer to the required questions?				
Does the methodology overlook any important factors?				
Are the results adequate to analyse the impact on the technical performance of the analysed technology? *				
Are the results adequate to analyse the impact on the economic performance of the analysed technology? *				
Are the results adequate to analyse the impact on the sustainability of the analysed technology? *				
Is the result unexpected or surprising in some way?				
Do you think that you would not have arrived to these conclusions without doing the forecast?				
Can you accept that the result is likely to be correct?				
Are the recommendations based on the output useful for decision-making?				
Did the forecast help to gain new insights about the environment of the company?				
Did the forecast help you to reduce uncertainties in projects or strategies...or to adapt to such uncertainties?				
Did the forecast lead to a better understanding the market (potential customers, customer needs, competition)?				
Does the forecast improve the capability of the company to accept or adopt alternative perspectives towards the future?				
Do you value the result in terms of usefulness and non-obviousness?				
Are the efforts needed for the whole methodology acceptable compared to the output it provides?				

* Depending on the objectives of the forecast, the appropriate question(s) should be selected